# A PROJECT REPORT

# ON

# YOUTUBE ADVIEW PREDICTION

Submitted in partial fulfilment for the requirement of the award of Internship

in

**Machine Learning**

*Submitted By*

**MALLIKARJUN TD**

**DAYANANDA SAGAR UNIVERSITY, BENGALURU**

# ACKNOWLEDGEMENT

My sincere gratitude and thanks towards my project paper guide Sanjoy Bhargab, Machine Learning Trainer. It was only with his backing and support that I could complete the report. He provided me all sorts of help and corrected me if ever seemed to make mistakes. I have no such words to express my gratitude and at last but not the least, I acknowledge my dearest parents for being such a nice source of encouragement and moral support that helped me tremendously in this aspect. I also declare to the best of my knowledge and belief that the Project Work has not been submitted anywhere else.

# INTRODUCTION

Youtube advertisers pay content creators based on adviews and clicks for the goods and services being marketed. They want to estimate the adview based on other metrics like comments, likes etc. The problem statement is therefore to train various regression models and choose the best one to predict the number of adviews. The data needs to be refined and cleaned before feeding in the algorithms for better results.

# Objective

To build a machine learning regression to predict youtube adview count based on other youtube metrics.

.

# Technology and Concepts

## Machine Learning

In classic terms, machine learning is a type of artificial intelligence that enables selflearning from data and then applies that learning without the need for human intervention.

## Linear Regression

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog).

There are two main types:

1. Simple regression.
2. Maltiple regression

## Support Vector Machine

“Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In

the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate

## Decission Tree

Decision tree analysis involves making a tree-shaped diagram to chart out a course of action or a statistical probability analysis. It is used to break down complex problems or branches. Each branch of the decision tree could be a possible outcome.

## Artificial Neural Network(ANN)

An artificial neural network (ANN) is the piece of a computing system designed to simulate the way the human brain analyzes and processes information. It is the foundation of artificial intelligence (AI) and solves problems that would prove impossible or difficult by human or statistical standards. ANNs have self-learning capabilities that enable them to produce better results as more data becomes available.

## Data Description

The file train.csv contains metrics and other details of about 15000 youtube videos. The metrics include number of views, likes, dislikes, comments and apart from that published date, duration and category are also included. The train.csv file also contains the metric number of adviews which is our target variable for prediction.

## Steps For adview prediction

1. Import the datasets and libraries, check shape and datatype.
2. Visualise the dataset using plotting using heatmaps and plots. You can study data distributions for each attribute as well.
3. Clean the dataset by removing missing values and other things.
4. Transform attributes into numerical values and other necessary transformations
5. Normalise your data and split the data into training, validation and test set in the appropriate ratio.
6. Use linear regression, Support Vector Regressor for training and get errors.
7. Use Decision Tree Regressor and Random Forest Regressors.
8. Build an artificial neural network and train it with different layers and hyperparameters. Experiment a little. Use keras.
9. Pick the best model based on error as well as generalisation.
10. Take the test dataset test.csv
11. Clean the test dataset by removing missing values

12. Remove unnecessary columns which has no impact to target variable

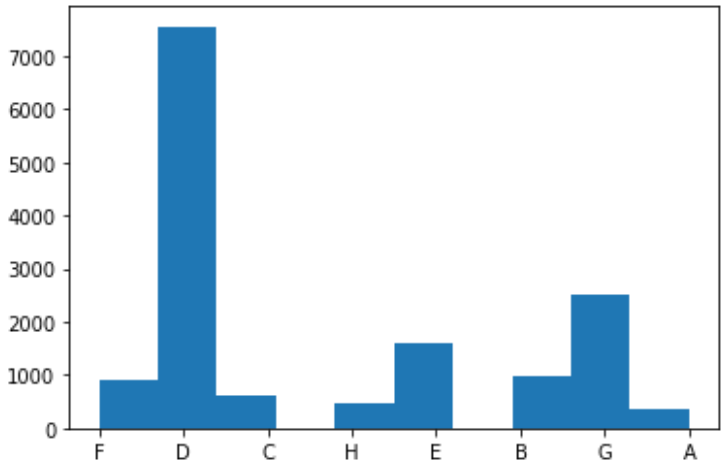
13. Transform the categorical attribute to numerical attribute.

14. Find prediction using the best algorithm

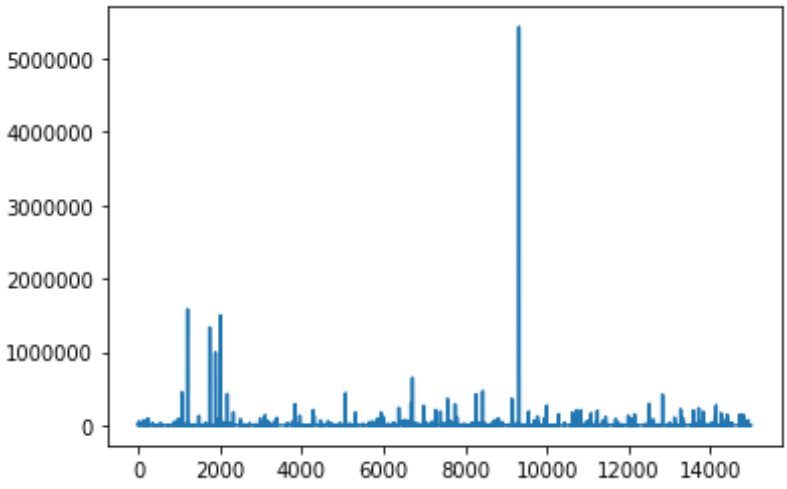
15. Save it into a new csv file by naming as Predictions_Submission.csv

## Visualization

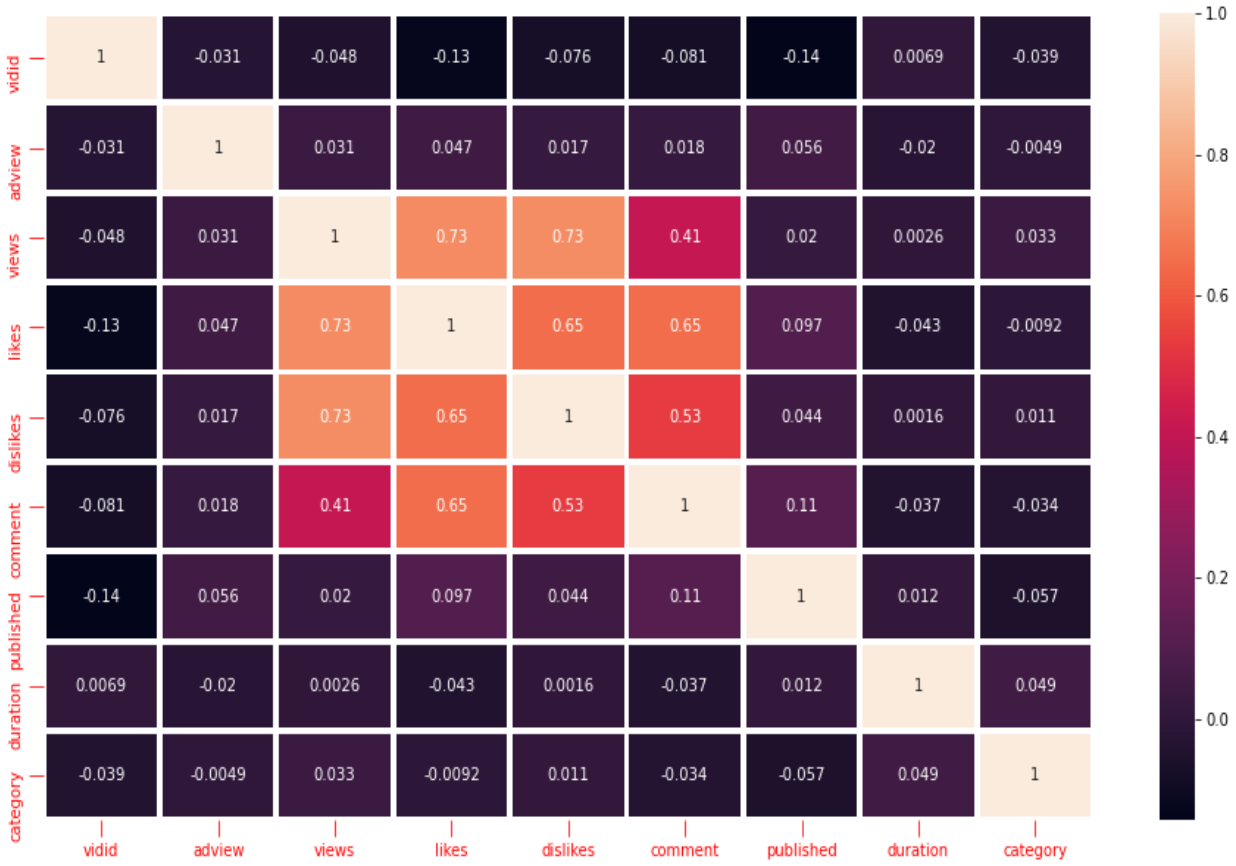
This is the histogram of “Category” column




This is the histogram of “adview” column



This is the heatmap which shows the co-relation of all columns with each other.



## Table:

| Algorithm | Linear Regression | Random forest | Decision tree | Support vector machine | ANN |
|---|---|---|---|---|---|
| Mean Absolute Error | 3707.378005824529 | 3274.6902966905504 | 3059.310792349727 | 3707.378005824529 | 3304.264894606637 |
| Mean Squared Error | 835663131.1210335 | 644433788.0361483 | 1226286165.4118853 | 835663131.1210335 | 829552666.7955565 |
| Root Mean Squared Error | 28907.838575573986 | **25385.700463766795** | 35018.3689713254 | 28907.838575573986 | 28801.955954335679 |
| | | | | | |

## Best Model

From the training dataset by applying all algorithms for train the model,we found that **"Random Forest Regressor"** algorithm has less root mean squared error as compared to othre algorithms.As we know model having **less root mean squared error** is more perfect.So here for prediction of test dataset we use **"Random Forest"** algorithm.

## Conclusions

We had a lot of different ideas for the project, but were maybe originally too ambitious for our goals. We were originally trying to predict the view count of advertisement. In this way we can predict the adview of an advertisement.We were hoping that. Some more things that we could have tried if we had more time would include.